Київський національний університет

імені Тараса Шевченка

Факультет комп'ютерних наук та кібернетики

Кафедра обчислювальної математики

Кваліфікаційна робота

на здобуття ступеня бакалавра

за спеціальністю 113 Прикладна математика

на тему:

АНАЛІЗ ЧАСОВИХ РЯДІВ ДЛЯ ПРОГНОЗУВАННЯ ПОДАТКОВИХ ДОХОДІВ ЗАСОБАМИ МАТЕМАТИЧНОЇ СТАТИСТИКИ ТА НЕЙРОМЕРЕЖ

Виконала студентка 4-го курсу

Липницька Поліна Денисівна _____

Науковий керівник:

Асистент кафедри обчислюваної математики, доктор філософії

Тимошенко Андрій Анатолійович _____

Засвідчую, що в цій роботі немає запозичень з праць інших авторів без відповідних посилань.

Студентка _____

Роботу розглянуто й допущено до захисту на засіданні кафедри обчислювальної математики

«_29_» __травня_____ 2023 р., протокол № _8_

Завідувач кафедри

С. І. Ляшко _____

Київ-2023

РЕФЕРАТ

Обсяг роботи 51 сторінки, 25 ілюстрацій, 17 джерел посилань, 1 додаток.

Ключові слова:

ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ, ПОДАТКОВІ ДОХОДИ, МЕТОД АВТОРЕГРЕСІЙНОГО ІНТЕГРОВНОГО РУХОМОГО СЕРЕДНЬОГО, ARIMA, SARIMA, НЕЙРОННІ МЕРЕЖІ, LSTM, РЕКУРЕНТНІ НЕЙРОННІ МЕРЕЖІ.

Об’єкт дослідження: методи математичної статистики та нейронних мереж для аналізу та прогнозування часових рядів.

Мета роботи: провести дослідження методів прогнозування та визначити кращі методи для аналізу макроекономічних показників.

Результати: було вивчено методи математичної статистики та нейромереж для аналізу та прогнозування податкових доходів країн, був проведений аналіз та обрані кращі методи прогнозування застосовні для даної сфери. Були визначені напрямки подальшого вдосконалення цих методів.

Результати роботи можуть бути використані для прогнозування даних податкових доходів різних країн та аналізу макроекономічного середовища.

ЗМІСТ

ВСТУП

З розвитком цифрових технологій та накопиченням великих даних аналіз часових рядів також дедалі більше удосконалюється. Окрім традиційних підходів математичної статистики подальший розвиток отримують методи штучного інтелекту. При розгляданні таких складних методів важлива не тільки їхня ефективність, безпомилковість, а й продуктивність, здатність виявляти нестандартні ситуації та незнайомі людині аномалії.

Аналіз часових рядів завжди мав високу актуальність. Він дозволяє покращити якість даних, побачити тенденції та на основі цього приймати обґрунтовані рішення, робити прогнози, виявляти аномалії та ризики.

На практиці це важливо для дослідження макроекономічних показників, таких, як податкові доходи. Податкові доходи, як основна частина державних доходів, мають усі показники відповідного предмета дослідження. У них спостерігається сезонність, циклічність, тренд, в той же час існують аномалії пов'язані з економічними кризами, епідеміями, соціальними потрясіннями. При цьому деякі, майже непомітні людині зміни, названі вченими «слабкими» сигналами, дозволяють передбачити певні особливості в часових рядах податкових доходів. У зв'язку з цим комбінація традиційних статистичних та нових (заснованих на нейронних мережах) методів аналізу часових рядів дуже важлива для даної макроекономічної сфери.

Метою роботи стало всебічне вивчення та вдосконалення методів аналізу часових рядів для формування ефективного комплексного інструменту (як комбінації даних методів), застосовного для прогнозування макроекономічних фінансових показників.

Об'єктами дослідження є методи математичної статистики та рекурентні нейронні мережі, що застосовні для аналізу часових рядів взагалі і зокрема макроекономічних даних, таких як податкові доходи країни, які були обрані як практичне застосування методів аналізу.

Методи та засоби дослідження: аналіз літературних джерел, порівняльний аналіз методів математичної статистики та штучного інтелекту, методи математичної дедукції та індукції.

Засобами розробки були мова програмування Python і її бібліотеки pandas, для зручної роботи з даними, numpy, sklearn, scipy для її аналізу, statsmodels, keras, tensorflow для побудови моделей та їх прогнозування та інші допоміжні бібліотеки.

Можливими сферами застосування є аналіз і прогнозування податкових доходів та інших економічних даних для покращення економіки та оцінки стану країни. Практичні результати роботи були використані Інститутом економіки промисловості Національної Академії наук України.

Результатом роботи стали практичні рекомендації щодо застосування сучасних методів аналізу часових рядів для макроекономічних даних. Джерелами даних послужили сайти Світового банку, Міжнародного валютного фонду та Міністерство фінансів України.

1 ТЕОРЕТИЧНИЙ ОГЛЯД МЕТОДІВ АНАЛІЗУ ЧАСОВИХ РЯДІВ

1.1 Визначення та характеристики аналізу часових рядів

З часом все більше зростає кількість даних, їх більше відслідковують, записують та отримують, дані часових рядів не є виключенням. Саме через це з'явилася й необхідність в їх аналізі. Ціллю аналізу часових рядів є пояснення та опис поведінки часових рядів, також для того щоб в подальшому можна було зробити прогноз поведінки цих даних в майбутньому.

Аналіз часових рядів є набором статистичних методів, що використовуються для аналізу даних які були зібрані за певний момент часу. Він є важливим для діагностики попередньої поведінки та прогнозування майбутньої. Аналіз часових рядів є застосовним до багатьох сфер життєдіяльності, як медицина, аналіз смертності чи може метеорологія і прогнозування погоди, чи навіть економіка і прогнозування зростання будь-яких показників.

Аналіз часових рядів може бути одновимірним або ж багатовимірним, коли з'являються додаткові незалежні змінні, які допомагають доповнити дані для кращого пояснення поведінки ряду. Багатовимірний аналіз є аналізом де два або більше залежні часові ряди моделюються разом, залежно один від одного, для отримання результатів у вигляді прогнозу кожного з аналізованих рядів.

Насамперед важливо уточнити, що таке часові ряди і навіщо взагалі потрібні дані в такому форматі.

Часовий ряд – це набір спостережень, кожне з яких було записане в визначений момент часу. Саме заданість у вигляді часової структури, де кожне спостереження залежить від попереднього є особливістю даних часових рядів. Рівномірна відстань між спостереженнями, що занотовані, є вирішальною для традиційних моделей. Часовий ряд може бути заданим як набір точок, що є проіндексованими або нанесеними на графік у часовому порядку.

Часові ряди можна розкласти на чотири важливі для аналізу компоненти:

1. Компонента тренду

   Тренд представляє собою довгострокову послідовність часового ряду, тобто описує довгострокові зміни.

   Тренд може бути зростаючим або спадаючим, це залежить від того чи демонструє часовий ряд довгострокове зростання чи зниження. Якщо ж часовий ряд не показує зростаючого або спадного тренду, його можна вважати стаціонарним за середнім значенням.
2. Циклічна компонента

   Циклічна компонента подібна до сезонної Іноді трапляється що лінія тренду проявляє постійне коливання, де вона рухається то вверх, тобто зростає, то вниз, тобто спадає. Цей неперервний рух вважається циклічним зразком. Тривалість цього зразка залежить від обраної галузі дослідження.
3. Сезонна компонента

   Сезонність – це явище, що виникає, коли часовий ряд має систематичну повторювану поведінку за певні періоди часу, циклічні зміни, що можуть бути щомісячними або квартальними кожного року.
4. Нерегулярна компонента

   Ця компонента часового ряду є завжди непередбачуваною та спонтанною, якщо казати простими словами, це те що залишається після прибирання попередніх компонент з даних. Також цю компоненту називають білим шумом або просто шум.

Часові ряди також можуть містити в собі відображення різних подій, що викликають структурні зміни у самих рядах. Такі події можуть бути як випадковими, такими як землетруси чи цунамі або ж запланованими, як наприклад рекламні кампанії чи цінові зміни.

1.2 Методи аналізу часових рядів

Цей розділ присвячений методам аналізу часових рядів, для правильного та ефективного застосування методів аналізу, важливо розуміти основні поняття математичної статистики, що є основними в аналізі та прогнозування часових рядів та вже після цього можна розглянути основні методи.

1.2.1 Основні теоретичні відомості

Послідовність випадкових величин $\{Y_t: t = 0, \pm1, \pm2, \pm3, \dots\}$ будемо називати *стохастичним процесом*. Повну ймовірнісну структуру такого процесу визначає набір розподілів всіх скінченних наборів Y.

Математичне сподівання дискретної випадкової величини є сумою помножених між собою значень випадкової величини і відповідних ймовірностей. Для випадкових неперервних величин визначається так:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx,$$

де $X -$ неперервна випадкова величина з функцією щільності ймовірності $f(x)$.

Дисперсія вимірює степінь розподілу числових значень випадкової величини, дисперсія випадкової величини X визначається так:

$$Var(X) = E\left[\left(X - E(X)\right)^2\right].$$

Коваріація X та Y визначається як

$$Cov(X, Y) = E[(X - \mu X)(Y - \mu Y)],$$

а *коефіцієнт кореляції* Xта Y:

$$\rho = Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Для стохастичного процесу $\{Y_t: t = 0, \pm1, \pm2, \pm3, \dots\}$ *функція середнього* визначається як

$$\mu_t = E(Y_t)$$

для $t = 0, \pm 1, \pm 2, \ldots$

Це означає, що μ_t є очікуваною величиною процесу в час t. Загалом, μ_t може різнитися для кожного моменту часу t.

Функція автоковаріації, $\gamma_{t,s}$, визначається як

$$\gamma_{t,s} = Cov(Y_t, Y_s)$$

для $t, s = 0, \pm 1, \pm 2, \ldots$

де

$$Cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t, Y_s) - \mu_t \mu_s.$$

Функція автокореляції, $\rho_{t,s}$, визначається як

$$\rho_{t,s} = Corr(Y_t, Y_s)$$

для $t, s = 0, \pm 1, \pm 2, \ldots$

де

$$Corr(Y_t, Y_s) = \frac{Corr(Y_t, Y_s)}{\sqrt{Var(Y_t)Var(Y_s)}} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}.$$

Стаціонарність є важливою характеристикою часових рядів. Часовий ряд називають стаціонарним, якщо його статистичні характеристики не змінюються з плином часу. Фактично це означає, що середнє значення та дисперсія мають постійні значення, а коваріація не залежна від часу, в деякому розумінні закони ймовірності не змінні з часом, тобто знаходиться в статистичній рівновазі.

Формальне визначення виглядає так:

Випадковий процес $\{Y_t\}$ називається *стаціонарним* тоді і лише тоді, коли для всіх цілих r, sта t виконується наступне:

- $E(Y_t) = \mu const;$
- $Var(Y|\quad|t) < \infty;$

- $\gamma_Y(t,s) = \gamma_Y(t+r, s+r)$.

Також варто було б розглянути поняття – випадкові блукання. Одновимірне дискретне випадкове блукання з дискретним часом задається:

$$Y_n = Y_0 + \sum_{i=1}^{n} X_i,$$

де $\{Y_n\} -$ випадковий процес, Y_0 – початковий стан,

а $X_i = \begin{cases} 1, p_i \\ -1, q_i \equiv 1 - p_i \end{cases}, 0 < p_i < 1, i \in N.$

Повернемося до нерегулярної компоненти, розглянутої в першому розділі, як залишок від часових рядів. Так названий білий шум або просто шум є важливим прикладом стаціонарного процесу, так як за допомогою білого шуму можна побудувати багато визначних та корисних процесів. що визначається як послідовність незалежних, однаково розподілених випадкових величин $\{e_t\}$. Середнє значення та варіативність білого шуму не змінні з часом, що дає змогу що це стаціонарний процес.

Термін білий шум виникає через таку аналогію, що як біле світло заходить однаково так і частоти при частотному аналізі. Зазвичай вважається, що процес білого шуму має середнє нульове значення і позначається $Var(e_t)$ як σ_e^2.

1.2.2 Методи математичної статистики для аналізу часових рядів

Математична статистика дуже давня наука, яку розвивало багато вчених та вивчало й продовжує вивчати багато людей. Тому методи математичної статистики в аналізі часових рядів є достатньо випробуваними і тому числі через поширене використання в багатьох сферах. Існує велика кількість розроблених

методів, що мають різні характеристики та призначення. Вибір та використання того чи іншого методу залежить від багатьох характеристик, ними можуть бути стаціонарність ряду, присутність тренду або сезонності. Також моделі можна розрізняти за простотою чи точністю при використанні. Далі буде розглянуто базові моделі та навіщо вони потрібні при аналізі та прогнозуванні часових рядів та більш складні та універсальні моделі.

Наївні моделі

Наївні моделі – найпростіші з моделей аналізу та прогнозування часових рядів. Їх використання можна звести до проблемних рядів чи рядів з дуже малою кількістю даних, через те що вони не проявляють великої точності та ігнорують всю інформацію про часовий ряд окрім останнього спостереження. Частіше вони використовуються як “база” для оцінки продуктивності тої чи іншої більш складної моделі, але й в деяких не надто складних випадках використання тільки наївної моделі для прогнозування буде слушним. За цією моделлю вважається, що найкращий показник того що буде завтра є те що було сьогодні.При аналізі часових рядів наївна модель часто розглядається як модель випадкового блукання.

Математично це можна подати в такому вигляді:

$$Y_t = Y_{t-1} + \varepsilon_t,$$

де Y_t та Y_{t-1} −прогрозоване в момент t та фактичне в момент $t-1$ значення та ε_t − помилка, тобто білий шум в момент t.

Перейдемо до розгляду моделей для стаціонарних часових рядів. Почнемо з розгляду загальних лінійних процесів.

Маємо $\{Y_t\}$ як спостережуваний часовий ряд та $\{e_t\}$ − неспостережуваний ряд білого шуму. Представити $\{Y_t\}$ можна як зважену лінійну комбінацію:

$$Y_t = e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots \quad (1)$$

Маємо нескінчений ряд, для того щоб задати його так щоб він мав математичний сенс, накладемо деякі умови на ψ-ваги. Припустимо:

$$\sum_{i=1}^{\infty} \psi_i^2 < \infty. (2)$$

Також варто зазначити, що так як $\{e_t\}$ є неспостережуваним, відсутність загальності рівняння (2) не втрачається, можна припустити що $e_t = 1$, тобто $\psi_0 = 1$. Важливим прикладом буде випадок, коли ψ утворюють експоненціальну затухаючу послідовність

$$\psi_j = \phi^j,$$

де $\phi -$ число між -1та $+ 1$.

Далі можна переписати:

$$Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots$$

Наприклад,

$$E(Y_t) = E(e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots) = 0$$

Так, що $\{Y_t\}$ має постійне середнє значення рівне нулю.

Матимемо:

$$Var(Y_t) = Var(e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots) =$$

$$Var(e_t) + \phi^2 Var(e_{t-1}) + \phi^4 Var(e_{t-2}) + \cdots =$$

$$\sigma_e^2(1 + \phi^2 + \phi^4 + \cdots) =$$

шляхом складання геометричної прогресії

$$\frac{\sigma_e^2}{1-\phi^2}.$$

Також,

$$Cov(Y_t, Y_{t-1}) = Cov(e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots, e_{t-1} + \phi e_{t-2} + \phi^2 e_{t-3} + \cdots) =$$

$$Cov(\phi e_{t-1}, e_{t-1}) + Cov(\phi^2 e_{t-2}, \phi e_{t-2}) + \cdots =$$

$$\phi\sigma_e^2 + \phi^3\sigma_e^2 + \phi^5\sigma_e^2 + \cdots =$$

$$\phi\sigma_e^2(1 + \phi^2 + \phi^4 + \cdots) =$$

$$\frac{\phi\sigma_e^2}{1-\phi^2}.$$

Таким чином

$$Corr(Y_t, Y_{t-1}) = \frac{\left[\frac{\phi\sigma_e^2}{1-\phi^2}\right]}{\left[\frac{\sigma_e^2}{1-\phi^2}\right]} = \phi.$$

Аналогічно, ми можемо знайти $Corr(Y_t, Y_{t-k}) = \frac{\phi^k\sigma_e^2}{1-\phi^2}$ та також

$$Corr(Y_t, Y_{t-k}) = \phi^k.$$

Важливо, що процес який ми визначили є стаціонарним – структура автоковаріації залежить лише від затримки в часі, а не від абсолютного часу. Для лінійного процесу (1) подібні обчислення дають:

$$E(Y_t) = 0$$

$$\gamma_k = Cov(Y_t, Y_{t-k}) = \sigma_e^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k}, k \geq 0 (3)$$

з $\psi_0 = 1$.

Процес з ненульовим середнім значенням μ можна отримати, додавши μ до правої частини рівняння (1). Так як середне не впливає коваріаційні властивості процесу, ми припускаємо нульове середнє значення до початку побудови моделей на основі даних.

Модель рухомого середнього (MA)

Моделі рухомого середнього скорочено записані як MA(moving average) були вперше розглянуті Слуцьким у 1927р. і Вольдом у 1938р.

У випадку, коли лише скінчена кількість вагових коефіцієнтів ψ не дорівнює нулю, ми маємо так званий процес рухомого середнього. В цьому випадку, ми змінюємо трохи нотацію і записуємо

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q}.$$

Будемо називати цю послідовність рухомим середнім порядку q і позначається коротким записом $MA(q)$.

Термін "рухоме середнє" має таку назву через спосіб його отримання:

Y_t отримують після застосування ваг $1, -\theta_1, -\theta_2, \ldots, -\theta_q$ до змінних $e_t, e_{t-1}, e_{t-2}, \ldots, e_{t-q}$, а далі переміщують ваги і застосовують їх до $e_{t+1}, e_t, e_{t-1}, \ldots, e_{t-q+1}$ для того щоб отримати Y_{t+1} і само так для наступних.

Рухоме середнє першого порядку (MA(1))

Після того як ми розглянули модель рухомого середнього, ми можемо перейти до розгляду специфікацій моделі і першою буде більш проста модель рухомого середнього першого порядку яку позначають як $MA(1)$.

Модель має вигляд $Y_t = e_t - \theta e_{t-1}$, оскільки залишається тільки одне значення θ, можна викинути зайвий запис 1.

Знаємо, що $E(Y_t) = 0$та$Var(Y_t) = \sigma_e^2(1 + \theta^2)$, тому

$$Cov(Y_t, Y_{t-1}) = Cov(e_t - \theta e_{t-1}, e_{t-1} - \theta e_{t-2}) =$$

$$Cov(-\theta e_{t-1}, e_{t-1}) = -\theta\sigma_e^2$$

та

$$Cov(Y_t, Y_{t-2}) = Cov(e_t - \theta e_{t-1}, e_{t-2} - \theta e_{t-3}) = 0$$

оскільки немає елементів eз спільними підписами між Y_tтаY_{t-2},

$$Cov(Y_t, Y_{t-k}) = 0 для k \geq 2,$$

що означає що процес не має кореляції поза затримкою 1.

Підсумовуючи вище вказане, маємо для моделі $MA(1)$ $Y_t = e_t - \theta e_{t-1}$

$$\begin{cases} E(Y_t) = 0 \\ \gamma_0 = Var(Y_t) = \sigma_e^2(1 + \theta^2) \\ \gamma_1 = -\theta\sigma_e^2 \\ \rho_1 = \dfrac{-\theta}{1 + \theta^2} \\ \gamma_k = \rho_k = 0 для k \geq 2 \end{cases}$$

Рухоме середнє другого порядку (MA(2))

Модель рухомого середнього першого порядку має вигляд $Y_t = e_t - \theta_1 e_{t-1}$, відповідно модель другого порядку буде мати вигляд: $Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$.

Тут

$$\gamma_0 = Var(Y_t) = Var(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}) = \sigma_e^2\left(1 + {\theta_1}^2 + {\theta_2}^2\right)$$

$$\gamma_1 = Cov(Y_t, Y_{t-1}) = Cov(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-1} - \theta_1 e_{t-2} - \theta_2 e_{t-3}) =$$

$$Cov$$

$$[-\theta_1 + (-\theta_1)(-\theta_2)]\sigma_e^2 =$$

$$(-\theta_1 + \theta_1\theta_2)\sigma_e^2$$

та

$$\gamma_2 = Cov(Y_t, Y_{t-2}) = Cov(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-2} - \theta_1 e_{t-3} - \theta_2 e_{t-4}) =$$

$$Cov(-\theta_2 e_{t-2}, e_{t-2}) =$$

$$-\theta_2\sigma_e^2$$

Маємо для другого порядку методу:

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{(-\theta_1 + \theta_1\theta_2)\sigma_e^2}{\sigma_e^2\left(1 + {\theta_1}^2 + {\theta_2}^2\right)} = \frac{-\theta_1 + \theta_1\theta_2}{1 + {\theta_1}^2 + {\theta_2}^2}$$

$$\rho_2 = \frac{\gamma_2}{\gamma_0} = \frac{-\theta_2\sigma_e^2}{\sigma_e^2\left(1 + {\theta_1}^2 + {\theta_2}^2\right)} = \frac{-\theta_2}{1 + {\theta_1}^2 + {\theta_2}^2}$$

$$\rho_k = 0, \text{для} k = 3,4, \ldots$$

Можна підсумувати попередньо наведені моделі і записати *загальний вигляд моделі рухомого середнього для порядку q – MA(q):*

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q},$$

маємо за тим самим принципом:

$$\rho_k = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q}{1 + {\theta_1}^2 + {\theta_2}^2 + \cdots + {\theta_q}^2} \\ 0, \text{для} k > q \end{cases},$$

для $k = 1,2, \ldots, q$.

Модель авторегресії (AR)

Модель авторегресії заснована на авторегресійних процесах. Авторегресійні процеси з самої назви є процесами що є регресією на самі себе. Одним з перших дослідження в цій галузі проводив Йуль у 1926 році.

Запишемо рівняння для авторегресійного процесу порядку p:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t (4)$$

Авторегресія першого порядку (AR(1))

Якщо ми маємо стаціонарний ряд та він задовольняє:

$$Y_t = \phi Y_{t-1} + e_t (5)$$

Розглянемо рівняння (4) та зробимо перетворення:

візьмемо дисперсію

$$\gamma_0 = \phi^2 \gamma_0 + \sigma_e^2,$$

розв’яжемо відносно γ_0:

$$\gamma_0 = \frac{\sigma_e^2}{1-\phi^2}.$$

Тепер розглянемо рівняння (4) та зробимо перетворення для нього (помножимо обидві його сторони на $Y_{t-k}, k = 1,2, \ldots$

$$Y_t Y_{t-k} = \phi Y_{t-k} Y_{t-1} + e_t Y_{t-k}$$

та візьмемо математичне сподівання:

$$E(Y_{t-k} Y_t) = \phi E(Y_{t-k} Y_{t-1}) + E(e_t Y_{t-k}).$$

Можна подати у вигляді

$$\gamma_k = \phi \gamma_{k-1} + E(e_t Y_{t-k}).$$

Так як ряд є стаціонарним з нульовим середнім значенням, e_t –незалежне від Y_{t-k}, тоді ми отримуємо

$$E(e_t Y_{t-k}) = E(e_t) E(Y_{t-k}) = 0$$

$$\gamma_k = \phi \gamma_{k-1} \text{для} k = 1,2,3, \ldots$$

Спробуємо підставити різні значення k:

$k = 1$:

$$\gamma_1 = \phi \gamma_0 = \frac{\phi \sigma_e^2}{1-\phi^2}.$$

$k = 2$:

$$\gamma_2 = \phi \gamma_1 = \frac{\phi^2 \sigma_e^2}{1-\phi^2}.$$

$k = 3$:

$$\gamma_3 = \phi\gamma_2 = \frac{\phi^3\sigma_e^2}{1-\phi^2}.$$

На цьому етапі можна побачити закономірність та записати в загальному випадку:

$$\gamma_k = \phi^k\gamma_{k-1} = \frac{\phi^k\sigma_e^2}{1-\phi^2}$$

або можна подати у вигляді

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \phi^k \text{для} k = 1,2,3, \ldots$$

Авторегресія другого порядку (AR(2))

Якщо для першого порядку ми записували рівняння як (5), то тут напишемо аналогічно тільки для другого порядку:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t.$$

Щоб поговорити про стаціонарність, можна ввести характеристичний поліном та відповідне характеристичне рівняння AR, нижче зазначено відповідно:

$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2$$

$$1 - \phi_1 x - \phi_2 x^2 = 0$$

Модель авторегресії рухомого середнього(ARMA)

Якщо ми маємо

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q},$$

то $\{Y_t\}$ – авторегресійно-рухомим середнім процесом порядків p і q, що записується скорочено $ARMA(p,q)$.

В міксованих моделях, таких як розглянута $ARMA$, ми припускаємо що немає спільних факторів в поліномах авторегресії та рухомого середнього, що означає що не можливо провести спрощення моделі, прибравши спільні фактори.

Тому важливо розглянути особливий випадок $ARMA(1,1)$, так як в цієї моделі ϕ відноситься до коефіцієнта авторегресії першого порядку, а θ - до коефіцієнта рухомого середнього першого порядку. Якщо $\theta = \phi$, ми можемо спростити модель до моделі нижчого порядку, залежно від значень цих параметрів.

Так, для моделі ARMA(1,1), умова $\theta \neq \phi$ важлива для забезпечення того, що модель не може бути додатково спрощена. Це допомагає уникнути надлишковості в моделі і забезпечує її статистичну ідентифікацію.

ARMA(1,1)

Запишемо рівняння, що визначає цю модель:

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1} (6)$$

Для того щоб отримати рівняння Йула-Волкера для авторегресійного процесу, запишемо:

$$E(e_t Y_t) = E[e_t(\phi Y_{t-1} + e_t - \theta e_{t-1})] = \sigma_e^2$$

та

$$E(e_{t-1} Y_t) = E[e_{t-1}(\phi Y_{t-1} + e_t - \theta e_{t-1})] =$$
$$\phi\sigma_e^2 - \theta\sigma_e^2$$
$$(\phi - \theta)\sigma_e^2.$$

Помножимо рівняння (6) на Y_{t-k}

$$Y_t Y_{t-k} = \phi Y_{t-k} Y_{t-1} + e_t Y_{t-k} - \theta e_{t-1} Y_{t-k}$$

і візьмемо очікування, отримуємо

$$\begin{cases} \gamma_0 = \phi\gamma_1 + [1 - \theta(\phi - \theta)]\sigma_e^2 \\ \gamma_1 = \phi\gamma_0 - \theta\sigma_e^2 \\ \gamma_k = \phi\gamma_{k-1} \text{для} k \geq 2 \end{cases}$$

Розв’язавши перші два рівняння:

$$\gamma_0 = \frac{1 - 2\phi\theta + \theta^2}{1 - \phi^2}\sigma_e^2$$

проста рекурсії дає:

$$\rho_k = \frac{(1 - \phi\theta)(\phi - \theta)}{1 - 2\phi\theta + \theta^2}\phi^{k-1} \text{для} k \geq 1$$

Загальний лінійний процес виглядає так:

$$Y_t = e_t + (\phi - \theta)\sum_{j=1}^{\infty}\phi^{j-1}e_{t-j},$$

$$\psi_j = (\phi - \theta)\phi^{j-1} \text{для} j \geq 1.$$

Модель авторегресії інтегрованого рухомого середнього (ARIMA)

Попередньо розглянуті моделі були застосовні для стаціонарних рядів, але зазвичай реальні дані не є стаціонарними, тому варто розглянути модель авторегресії інтегрованого рухомого середнього, що є застосовною для нестаціонарних даних. Вона є моделлю авторегресії рухомого середнього, але з додатковим параметром – (I) інтегрованість. Розглянемо, що означає цей параметр та саму модель нижче.

Взяття різниці для досягнення стаціонарності

Параметр інтегрованості(I) показує порядок різниці, що потрібна для того щоб перетворити часовий ряд з нестаціонарного у стаціонарний.

Покажемо яким чином відбувається процес взяття різниці:
Першою різницею буде

$$\nabla Y_t = Y_t - Y_{t-1},$$

буває що першої різниці може не вистачити, тобто ряд все ще не буде стаціонарним, тоді можна взяти другу різницю. Кількість взятих різниць буде відображатися в параметрі d моделі $ARIMA(p, d, q)$.

Сформуємо формальне визначення:

Часовий ряд $\{Y_t\}$ будемо називати моделлю *з інтегрованою авторегресійною рухомою середньою*, якщо d-та різниця $W_t = \nabla^d Y_t$ є стаціонарним ARMA процесом.
Якщо ж $\{W_t\}$відповідає $ARMA(p, q)$моделі, будемо казати, що $\{Y_t\}$ є $ARIMA(p, d, q)$процесом.

Модель сезонної авторегресії інтегрованого рухомого середнього (SARIMA)

Модель сезонної авторегресії інтегрованого рухомого середнього є розширенням попередньо розглянутої моделі авторегресії інтегрованого рухомого середнього, а саме додання сезонної компоненти для рядів, що проявляють сезонність.

Модель SARIMA позначають як SARIMA$(p, d, q)(P, D, Q)_m$,

де $p, d, q -$ кількість компонентів авторегресії, кількість різниць оригінального ряду та кількість компонент рухомого середнього відповідно,

а $m -$ кількість періодів у кожному сезоні. P, D, Q є компонентами авторегресії, різниць та рухомого середнього сезонної частини.

Отже, модель SARIMA включає в себе аналіз як загальних трендів у даних, так І сезонних змін. Це робить цей метод особливо корисним для часових рядів, як виявляють сезонні властивості.

1.2.3 Нейронні мережі та їх застосування у прогнозуванні

Нейронні мережі – це модель машинного навчання, що демонструє сильний розвиток в сучасні часи. Ця моделі отримала велику популярність через те, що була створена аналогічно до роботи людського мозку, вони імітують роботу центральної нервової системи живих організмів, мабуть, найкраще за попередні техніки та моделі. Вони є базисом глибокого навчання і впливовою галуззю штучного інтелекту, яка імітує принципи роботи людського мозку для розуміння і вивчення шаблонів у даних.

Основу нейронних мереж складає система так званих «зв’язків», які пов’язують між собою велику кількість шарів вузлів або «нейронів». Кожен нейрон обробляє вхідні дані і передає результат наступному шару нейронів.

Мережі в свою чергу поділяються на шари (рис.1):

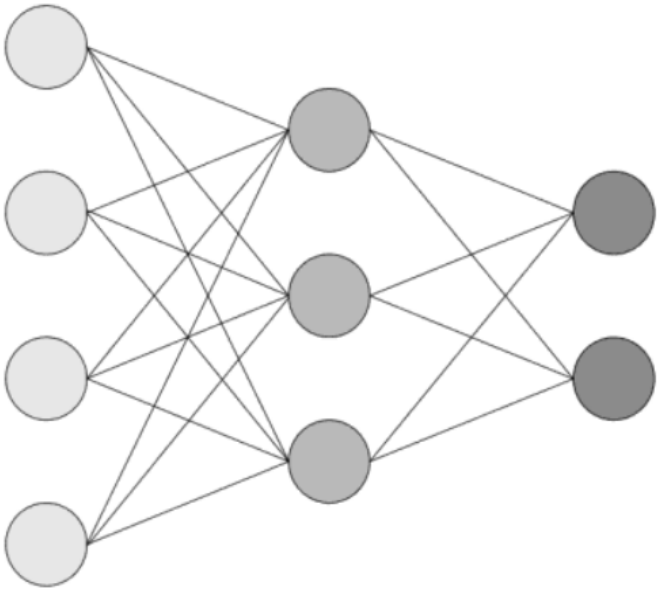

Рисунок 1 – Графічно представлені шари нейромереж

- **Вхідний шар** приймає значення вхідних даних.
- **Приховані шари** роблять на основі вхідних даних обчислення та спрямовують результати мережею далі.
- **Вихідний шар** видає кінцевий результат або прогноз.

Сила зв'язку між двома нейронами виражається вагою. Чим більше абсолютне значення ваги, тим більший вплив однієї одиниці на іншу.

- Позитивна вага виражає, що один нейрон чинить збудливий вплив на інший нейрон.
- Від’ємна вага означає, що вплив має гальмівний, тобто гальмівний характер.
- Нульова вага означає, що один нейрон в даний момент не впливає на інший нейрон.

Знання нейронної мережі зберігаються в її вагових коефіцієнтах.

У нейронних мережах навчання здебільшого визначається як зміна ваги між одиницями. Як саме відбувається зміна ваги, залежить від використовуваного правила навчання.

В кожному нейроні відбуваються обчислення, що є лінійною комбінацією вхідних даних. Використовуючи процес, відомий як зворотне розповсюдження помилки, нейронні мережі можуть вчитися, визначаючи помилку між їхніми прогнозами та дійсними даними, а потім відкореговуються ваги мережі, щоб зменшити цю помилку.

Варто зазначити, що нейронні мережі - дуже потужні і гнучкі моделі. Вони мають можливість апроксимувати велику кількість функцій та обробляти комплексні завдання, наприклад машинний переклад, розпізнавання зображень, гру в комп’ютерні ігри тощо. Але важливо розуміти, що вони потребують величезні об’єми даних, бувають складними в інтерпретуванні, а також схильні до перенавчання, якщо обробка не проходить належним чином.

З моменту свого виникнення, було розроблено декілька типів нейронних мереж для різних застосувань, включаючи штучні нейронні мережі (ANN), згорткові нейронні мережі (CNN), та рекурентні нейронні мережі (RNN).

Рекурентні нейронні мережі (RNN) - це вид нейронних мереж, які були спеціально розроблені для обробки послідовностей даних, таких як часові ряди або природна мова. Ключова особливість RNN - це їхня здатність "запам'ятовувати" інформацію з попередніх кроків, що робить їх ідеальними для

виконання завдань, де контекст із минулих вхідних даних важливий для прогнозування поточного чи майбутнього виходу. У традиційних нейронних мережах кожен вихід розглядається незалежно від інших.

Вихід в RNN на кожному етапі часу розглядається у виді функції не тільки поточного вводу, але і «стану» мережі на попередньому етапі. Цей стан є деякою формою «пам'яті», в якій зберігається інформація про попередні вводи.
На жаль, на прикладах багатьох спостережень дослідники виявили, що важко навчити RNN охоплювати довгострокові послідовності, оскільки градієнти мають тенденцію або зникати (так звана проблема «згасання градієнту»), або тенденцію вибуху.

Явище зникаючого градієнта полягає в тому, що важко охопити довгі послідовності через мультиплікативний градієнт, який може експоненційно збільшуватися або зменшуватися до кількості шарів. Тож це ускладнює метод оптимізації на основі градієнта не лише через варіації величин градієнта, а й через те, що довготривалі послідовності приховані ефектом короткочасних послідовностей.

Було два домінуючих підходи, якими керувалися багато дослідників, що намагалися зменшити негативний вплив цієї проблеми. Одним із таких підходів є розробка кращого алгоритму навчання, ніж простий стохастичний градієнтний спуск, наприклад, використовуючи дуже простий обрізаний градієнт, який обрізає норму вектора градієнта, або за допомогою методів другого порядку, які можуть бути менш чутливими, якщо другі похідні розвиваються так само, як і перші похідні.

Інший підхід, в якому ми більш зацікавлені, полягає в розробці більш складної функції активації, ніж звичайна, що складається з подальшого перетворення. Спроби в цьому напрямку призвели до функції активації, яка називається LSTM (Long Short-Term Memory або довга короткочасна пам'ять), а також GRU (Gated Recurrent Units або вентильні рекурентні вузли). Ці варіації

RNN дуже ефективно виконують завдання, які вимагають обробки складних послідовностей.

RNN мають вигляд повторювальної ланцюгової структури, де сам модуль повторення має простий вид, наприклад tanh. На відміну від цього в LSTM модуль повторення має більш комплексний вид. Цю різницю можна побачити на схемах наданих на рисунках 2.1 та 2.2.

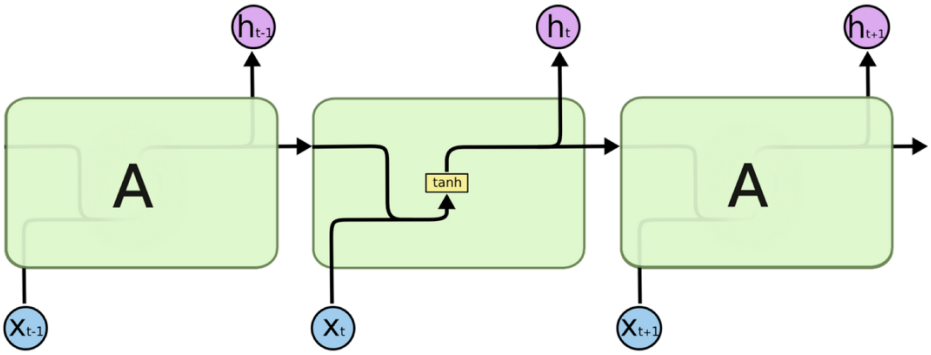

Рисунок 2.1 – Графічна схема класичної RNN

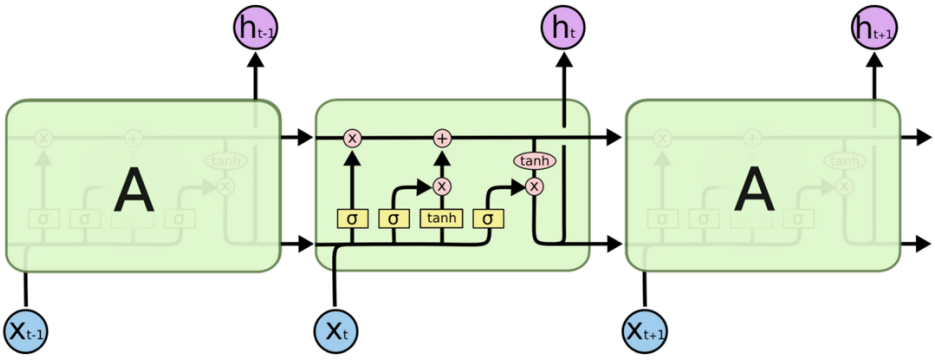

Рисунок 2.2 – Графічна схема варіації рекурентних нейромереж LSTM

В LSTM використовуються спеціальні блоки, які називаються "блоками пам'яті". Кожен такий блок має три важливі компоненти: вхідний вентиль, вентиль забуття і вихідний вентиль. Вхідний вентиль визначає, яка нова інформація буде збережена в стані пам'яті, вентиль забуття визначає яка інформація збережена раніше буде видалена і вихідний вентиль визначає яка інформація буде передана на вихід. Якщо блок LSTM інтуїтивно виявляє важливу функцію вхідної послідовності на ранній стадії, він розпізнає це і легко переносить цю

інформацію далі, таким чином фіксуючи потенційну залежність на великій відстані.

Математично компоненти блоку пам'яті, тобто вентилі обчислюються так:
Маємо f_t, i_t, o_t – відповідно вектори вентиля забуття, вентиля входу та вентиля виходу,

$$\begin{bmatrix} f_t \\ i_t \\ o_t \end{bmatrix} = \begin{bmatrix} \sigma\big(W_f x_t + U_f h_{t-1} + b_f\big) \\ \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ \sigma(W_o x_t + U_o h_{t-1} + b_o) \end{bmatrix},$$

а x_t, h_t позначають стан входу, стан виходу за час t, а b – вектор зміщення. У простому шарі нейронної мережі зміщення додається до зваженого сумарного входу для кожного нейрона перед застосуванням функції активації.

W, U є композиціями матриць ваги і уточнення, при цьому вони мають однакову глибину.
Застосовуючи стан входу x_t та попередні стани c_{t-1} та h_{t-1}, отримаємо новий стан комірки c_t, а також нове значення стану виходу h_t:

$$c_t = f_t \circ c_{t-1} + i_t \circ tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$h_t = o_t \circ tanh(c_t).$$

Також важливо зазначити, що функції активації, а саме сигмоїдна функція σ та гіперболічний тангенс tanh мають такий загальний вигляд:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \qquad tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

та мають такі графіки (рис.3).

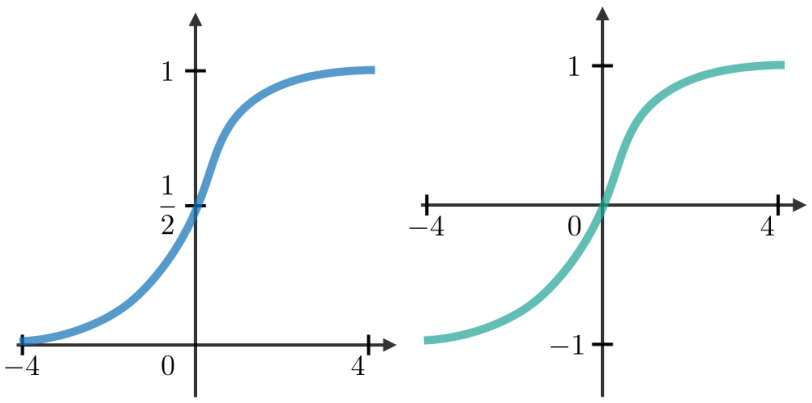

Рисунок 3

Ця структура дозволяє LSTM вчитися і запам'ятовувати довгострокові залежності в даних, що робить їх особливо корисними для аналізу часових рядів, мови, і інших даних з послідовностями.

2 МЕТОДОЛОГІЯ ДОСЛІДЖЕННЯ ЧАСОВИХ РЯДІВ З УРАХУВАННЯМ ОСОБЛИВОСТЕЙ ЕКОНОМІЧНИХ ДАНИХ

2.1 Вибір та опис даних податкових доходів

Було обрано 8 країн світу для аналізу, на рисунку 1 наведені графіки податкових доходів за період з 1972р. по 2021р. за даними [1] у відсотках від ВВП. Дані задані у вигляді часових рядів, якщо відфільтрувати дані за країнами, ми отримаємо вісім окремих часових рядів за кожною країною.

Об’єднавши ряди всіх країн та зобразивши їх на одному графіку (рис.4), отримаємо:

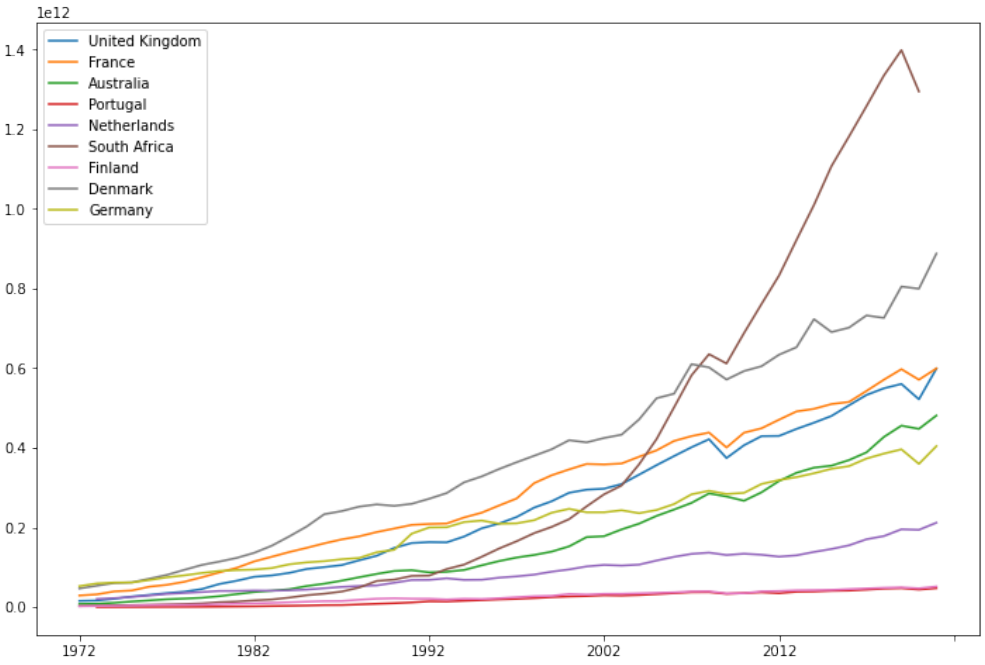

Рисунок 4 - Податкові доходи країн по роках

Як можна побачити з графіку, циклічність у всіх рядів є досить схожою. Це відбувається через пов’язаність всіх країн у світовій економіці. Зазвичай, вона виникає у зв'язку з тим, що світова економіка стає все більш взаємопов'язаною через глобалізацію.

Торгівля, інвестиції, міжнародні кредити та глобальні ланцюги поставок - все це фактори, які впливають на економічні цикли країн. Якщо в одній країні спостерігається економічне зростання або спад, це може вплинути на її торговельних партнерів і, відповідно, на їх податкові доходи. Це пояснює схожість циклічності у всіх рядів.

Спробуємо проаналізувати країни окремо, до прикладу візьмемо країну Німеччину зі спільного графіку і додатково візьмемо дані Німеччини поквартально за період з 2009р. по 2022р., дані взято з ресурсу [2].

2.2 Підготовка обраних даних до використання моделей

Для аналізу та прогнозування часових рядів вкрай необхідно обрати правильну модель. Вибір відповідної моделі є важливим так як впливає на точність прогнозування та коректність результатів. Існує багато різних моделей, деякі є більш спрощеними, деякі розроблені для даних що мають тренд або сезонність. Тому для того щоб обрати найкращу модель необхідно провести аналіз даних та оцінювання на тестовому наборі даних для вибору модель для подальшого прогнозування.

2.2.1 Моделі математичної статистики

Варто побудувати графік та зробити певний аналіз даного часового ряду. Розпочнемо з візуального аналізу: з огляду на графіки на рисунку 5.1 та рисунку 5.2, можна зробити висновок, що ряди не є стаціонарними, адже простежується

тренд та дисперсія і середнє значення є змінними з часом, на рис 5.2 можна побачити також сезонність.

Але перевіримо стаціонарність також за допомогою невеликого тесту: візьмемо середні значення і дисперсію на різних проміжках та проаналізуємо їх. На рисунку 6 наведено результати. Бачимо, що середні значення дуже відрізняються, а відповідно дисперсія також є різною на проміжках часових рядів, отже обидва часові ряди не є стаціонарними та будуть потребувати моделі для нестаціонарних рядів. Досягти стаціонарності можна шляхом диференціювання (процес взяття різниці між послідовним спостереженнями). Занотуємо все розглянуте для подальшого використання у прогнозуванні.

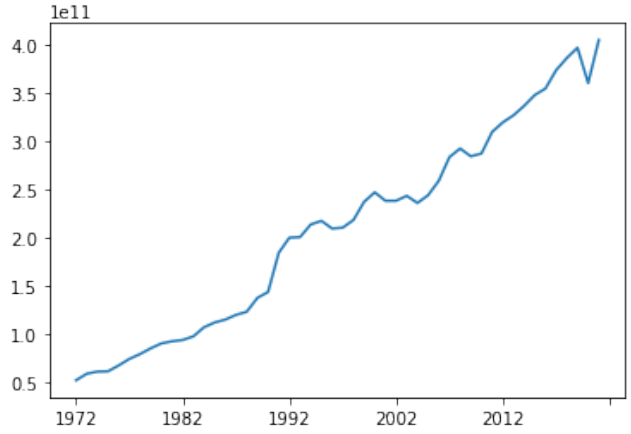

Рисунок 5.1 - Графічне зображення податкових доходів Німеччини по рокам

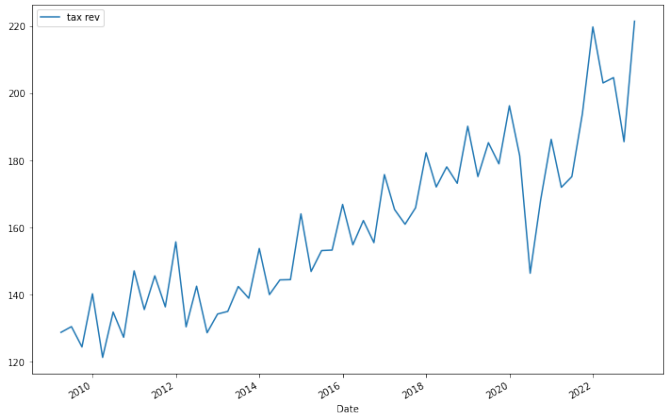

Рисунок 5.2 - Графічне зображення податкових доходів Німеччини поквартально

```
Середне значення перших 10 елементів: 72807452590.46031
Середне значення останніх 10 елементів: 170841249479.1214
Дисперсія_1: 1.9973857218812785e+20
Дисперсія_2: 8.517537862222221e+20
```

```
Середне значення перших 10 елементів: 133.5813
Середне значення останніх 10 елементів: 149.597775
Дисперсія_1: 608.1568377737015
Дисперсія_2: 371.6446666666667
```

Рисунок 6 - Результат роботи коду з перевіркою середніх значень та дисперсії

Для досягнення стаціонарності у часового ряду на рисунку 5.1 достатньо взяти одну різницю. У випадку з часовим рядом з рисунка 5.2. однієї та навіть двох різниць не було достатньо. Стаціонарність буде досягнута лише при 9 різниці, що є дуже великим значенням. Відбувається це через велику кількість

викидів, побудуємо діаграми розсіювання для обох випадків, рисунок 7.1 та 7.2. Також таке може відбуватися через складну сезонність або складний тренд.

На рисунку 7.2 можна побачити рівномірність, що допомагає досягти стаціонарності швидше, в той час як на рисунку 7.1 є велика кількість викидів, що не дозволяє швидко прийти до стаціонарності шляхом взяття різниць.

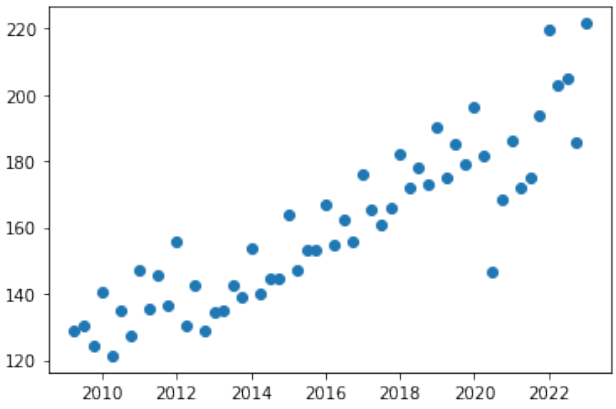

Рисунок 7.1 - Графічне зображення діаграм розсіювання податкових доходів Німеччини по рокам

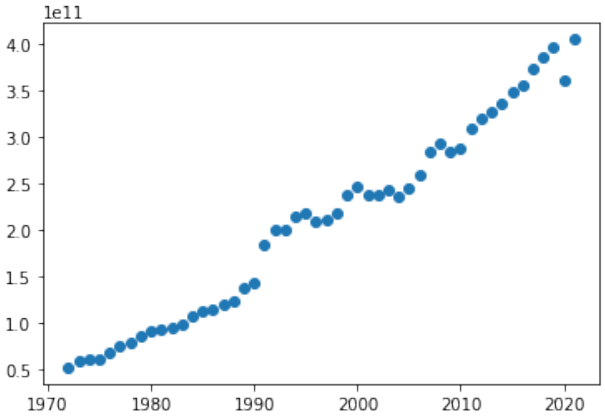

Рисунок 7.2 - Графічне зображення діаграми розсіювання податкових доходів Німеччини поквартально

Розглянемо ряди та впевнимося в присутності сезонності. У випадку щорічних даних тренд можна побачити під час візуального огляду, у квартальних даних достатньо очевидно, що буде. мати місце сезонність, щоб впевнитись у цьому побудуємо графіки декомпозиції даних. Декомпозицію проведемо за допомогою адитивної моделі, що полягає у тому що ряд розкладається на суму тренду, сезонності та залишків. Адитивна модель на відміну від мультиплікативної підходить більше для даних сезонність яких має приблизно рівні коливання за трендом, що можна побачити на квартальних даних. В той час як для щорічних даних не надто важливий вибір моделі, адже сезонності там немає.

Дійсно, розглядаючи рисунки 8.1 та 8.2 можна впевнитись у відсутності сезонності у щорічних даних та побачити чітку сезонність у квартальних даних, що у випадку податкових походів є досить логічним та пояснювальним.

Так як у цій роботі буде розглянуто моделі ARIMA та її сезонне доповнення SARIMA як моделі математичної статистики, обираємо модель авторегресійної інтегрованої рухомої середньої для щорічних даних, адже вони не демонструють сезонності та мають тренд та модель сезонної авторегресійної інтегрованої рухомої середньої для квартальних даних, вони мають чітку сезонність та тренд.

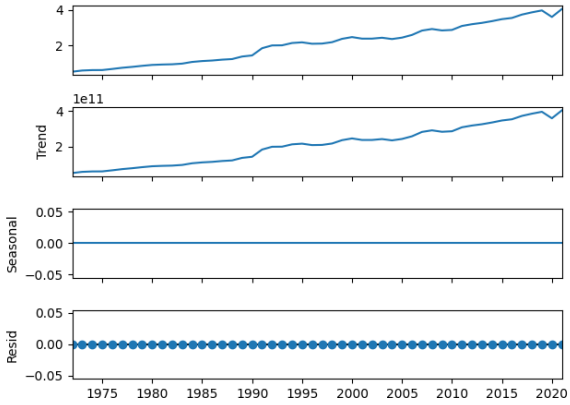

Рисунок 8.1 - Графічне зображення декомпозиції податкових доходів Німеччини по рокам

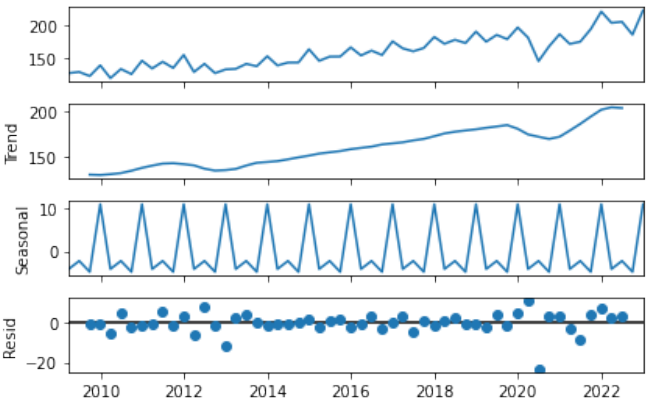

Рисунок 8.2 - Графічне зображення декомпозиції податкових доходів Німеччини поквартально

2.2.2 Методи нейронних мереж

За допомогою нейронних мереж можна побудувати прогноз для часових рядів наведених вище на рисунках 5.1 та 5.2. Для прогнозування даних зробимо деякі перетворення.

Для побудови моделі проведемо масштабування даних методом мінімаксу як підготовку даних. Основна мета масштабування - перетворити значення з

одного діапазону в інший, зберігаючи відносні пропорції між значеннями. Це досягається шляхом віднімання мінімального значення та поділу на різницю максимального та мінімального значень. Мінімаксальне масштабування проводиться для нормалізації даних, що в свою чергу покращує якість пронозів та прискорює навчання моделі.

Розділимо дані на тренувальні та тестові у такій самій пропорції як у попередньому аналізу. Виділяється вісімдесят відсотків на тренувальні дані та залишки на тестування.

Для тренування поділимо дані на послідовності, де кожна послідовність містить вказану кількість даних, оберемо 5. Тренувальні і тестові послідовності сформовані за принципом де усі окрім останнього значення кожної послідовності заносяться у змінні x_train та x_test, в той час як останні значення записуються у y_train та y_test. Це зроблено для того щоб модель могла використовувати x_train для прогнозування наступного значення, що розташоване в y_train, теж саме для тестування.

2.3 Побудова моделей на основі аналізу даних

З попереднього аналізу були обрані методи математичної статистики такі як: модель авторегресійної інтегрованої рухомої середньої для щорічних даних та модель сезонної авторегресійної інтегрованої рухомої середньої для квартальних даних. Для нейронних мереж був обраний метод LSTM, що є варіацією рекурентних нейронних мереж.

2.3.1 Побудова моделей математичної статистики

Як відомо з теоретичної частини, для того щоб побудувати модель ARIMA потрібно обрати параметри. Ця модель має три параметри – p,q,d, де p відповідає за порядок авторегресії, q за порядок рухомого середнього та d за порядок диференціювання, тобто кількість взятих різниць.

Побудуємо графіки автокореляційної та частково автокореляційної функцій(рис. 9.1 та 9.2).

З огляду на графіки можна побачити що значимі лаги(значення з часового ряду в попередній крок) на рисунку 9.1 спостерігаються до 4 та на рисунку 9.2 значним є тільки перший. Параметр d нам відомий, адже раніше ми вже перевірили потрібну кількість різниць для стаціонарності.

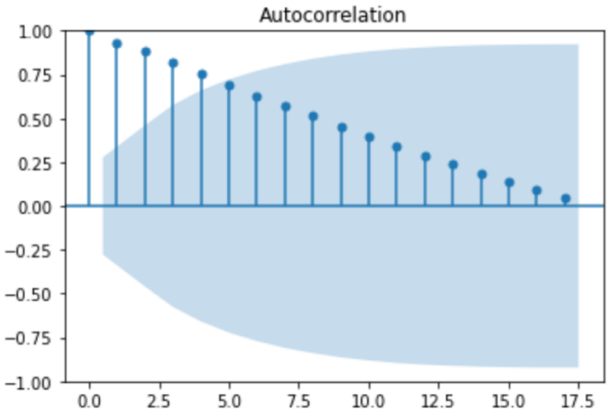

Рисунок 9.1 – Графік автокореляційної функції податкових доходів Німеччини по рокам

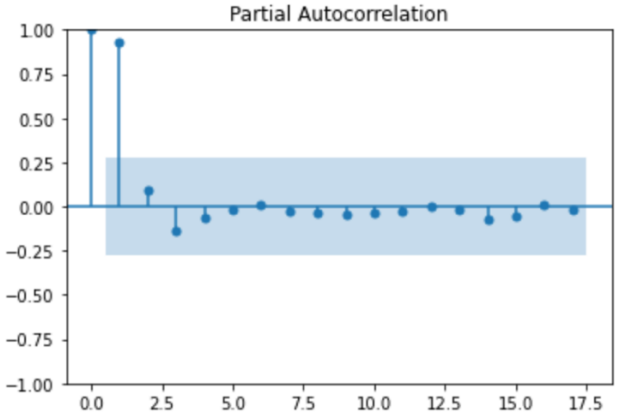

Рисунок 9.2 – Графік частково автокореляційної функції для податкових доходів Німеччини по рокам

Проведемо такий самий аналіз для квартальних даних. З аналізу вище відомо, що ряд має сезонну компоненту, тому проведемо побудову моделі сезонної авторегресійної інтегрованої рухомої середньої (SARIMA). Дана модель також має такі самі параметри, так як є розширенням моделі ARIMA, але має сезонні компоненти для прогнозування сезонності. У попередніх пунктах було

проведено спробу досягти стаціонарності за допомогою різниць, але для її досягнення потрібно було взяти завеликий порядок, що призведе до перенавчання та ускладнення моделі, а відповідно і до гірших результатів. Тому спробуємо не робити диференціювання взагалі і позначимо параметр d як нуль.

Параметри що відповідають за порядок авторегресії та за порядок рухомого середнього визначимо так само як раніше. Розглядаємо графіки автокореляційної та частково автокореляційної функцій і так само підбираємо параметри для моделі. Модель має сезонні та не сезонні компоненти в такому порядку $SARIMA(p,d,q)(P,D,Q)_m$. Сезонна частина є стаціонарною і тому параметр диференціювання для сезонних компонент буде 0, інші сезонні параметри беремо з аналізу графіків з додатків А.4 та А.5. За аналізом графіків наведених на рисунках 10.1 та 10.2 можна побудувати модель з параметрами $SARIMA(3,0,3)(14,0,12)_4$.

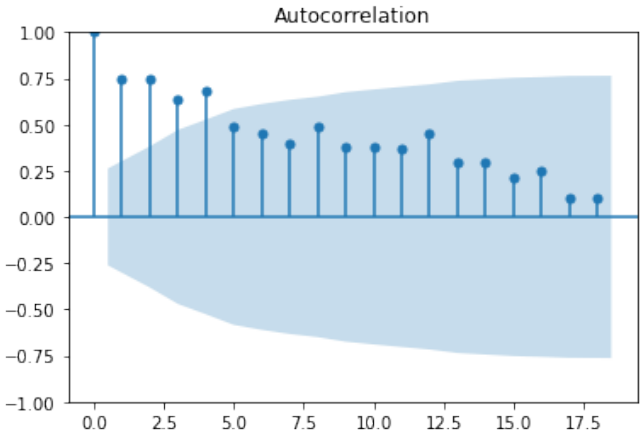

Рисунок 10.1 – Графік автокореляційної функції податкових доходів Німеччини по кварталам

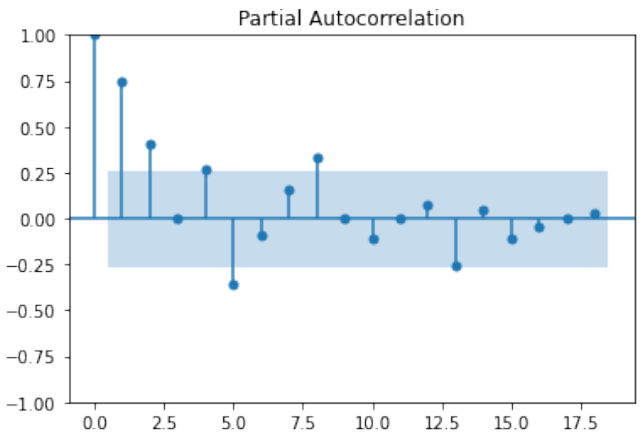

Рисунок 10.2 – Графік частково автокореляційної функції для податкових доходів Німеччини по кварталам

2.3.2 Побудова моделі за допомогою нейронних мереж

Для побудови моделі LSTM(модель довгої короткочасної пам’яті) та роботи з шарами будемо використовувати *tensorflow* та *keras*, бібліотеки для мови програмування Python для глибинного машинного навчання та тренування глибоких нейронних мереж.

Специфічна структура LSTM дозволяє вона запам'ятовувати або забувати інформацію протягом тривалого періоду часу, що робить її особливо ефективною для завдань, які вимагають врахування контексту з минулого для розуміння поточного стану.

В цьому випадку буде використана двонаправлена модель довгої короткочасної пам’яті, так як двонаправлена модель використовує минулі та майбутні точки, створюючи дві копії LSTM блоків, один з яких проходить вхідні дані в прямому напрямку (від минулого до майбутнього), а інший - в зворотному (від майбутнього до минулого).

Будуємо модель довгої короткочасної пам’яті LSTM. Створюємо модель із вхідним шаром, з двома двонаправленими шарами і вихідний шар. Також застосовується dropout шар для запобігання перенавчанню. Додається Dense шар з одним вихідним нейроном для прогнозування. Далі проведемо навчання моделі з використанням тренувальних даних.

На рисунку 11.1 та 11.2 зображені втрати тренування та втрати валідації для щорічних та квартальних даних відповідно, на осі х-ів маємо епохи, а на осі у-ків значення втрат. Втрати валідації вимірюють рівень помилок моделі на наборі валідаційних даних, використовуються для оцінки того, наскільки добре модель узгоджується з даними, які вона не бачила під час тренування.

Втрати валідації під час тренування можуть зрости, це буде означати перенавчання моделі. Тому при підборі параметрів варто звертати увагу на графіки втрат, за допомогою них можна підібрати ідеальні параметри для прогнозу.

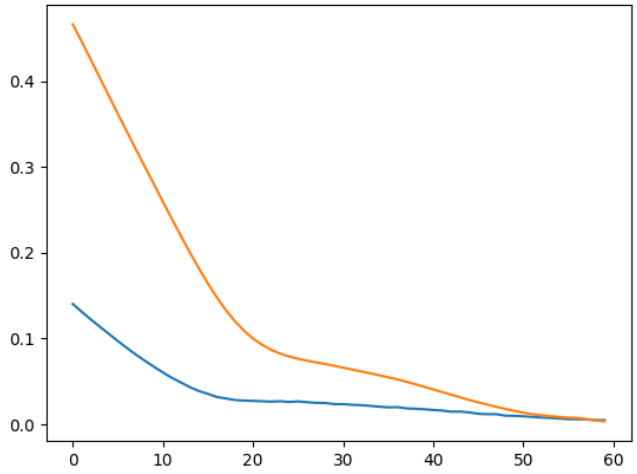

Рисунок 11.1 – Графік втрат моделі для щорічних даних

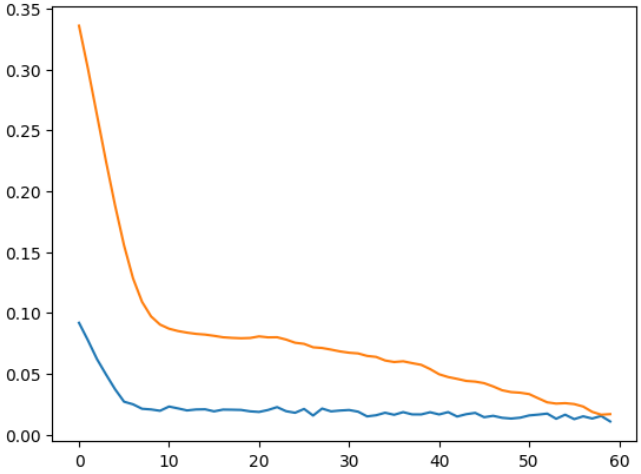

Рисунок 11.2 – Графік втрат моделі для квартальних даних

3 РЕЗУЛЬТАТИ ПРОГНОЗУВАННЯ ПОДАТКОВИХ ДОХОДІВ

3.1 Оцінка моделей на тренувальному наборі даних

Спробуємо побудувати модель з параметрами ARIMA(1,1,4). Зробимо перевірку моделі на тестових даних, рисунок 12, також були оцінені параметри моделі додатково, параметри зазначені у додатку А.1.

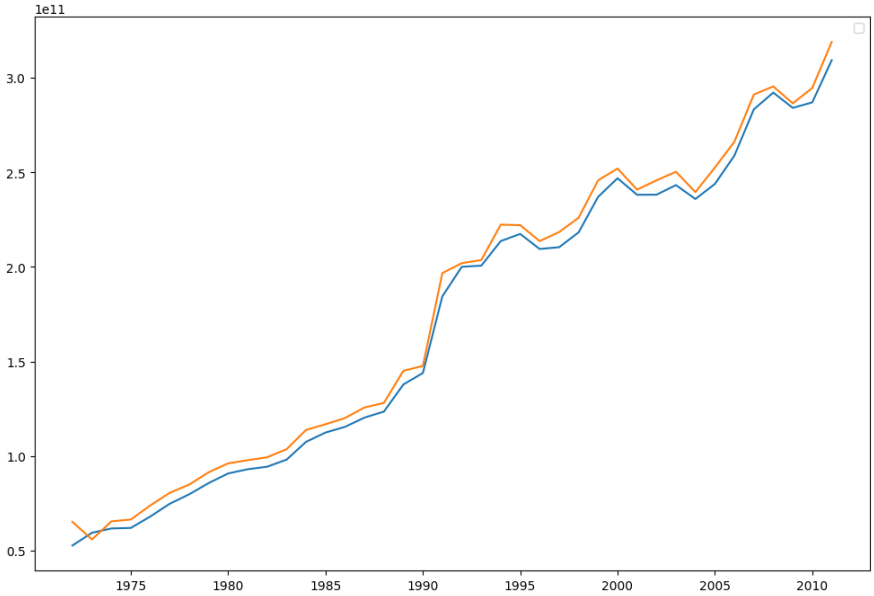

Рисунок 12 – Результат роботи моделі на тестових даних

Проведемо оцінку адекватності моделі, графічно можна зазначити що модель добре прогнозує дані, також перевіримо залишки. У додатку А.2 маємо результати тестів Льюнга-Бокса та Дікі-Фуллера. За результатами можна зазначити, що нульова гіпотеза про відсутність автокореляції залишків не відхиляється та залишки є стаціонарним рядом, що дозволяє нам вважати цю модель підходящою для прогнозування.

За допомогою зазначеної у другому розділі моделі зробимо перевірку моделі на тестових даних, рисунок 13 ($SARIMA(3,0,3)(14,0,12)_4$. Перші кілька точок часового ряду можуть бути проблематичними, через те що модель вчиться на основі попередніх даних, так як в перших точках не має достатньої кількості попередніх спостережень, на які модель могла б опиратися. Сезонна компонента вимагає наявності достатньої кількості спостережень для кожного сезонного періоду.

З часом модель навчається і прогноз стає більш стабільним, поглянемо також на тести у додатку А.3. Аналогічно до попередніх випадків маємо що у залишках немає автокореляції та ряд є стаціонарним.

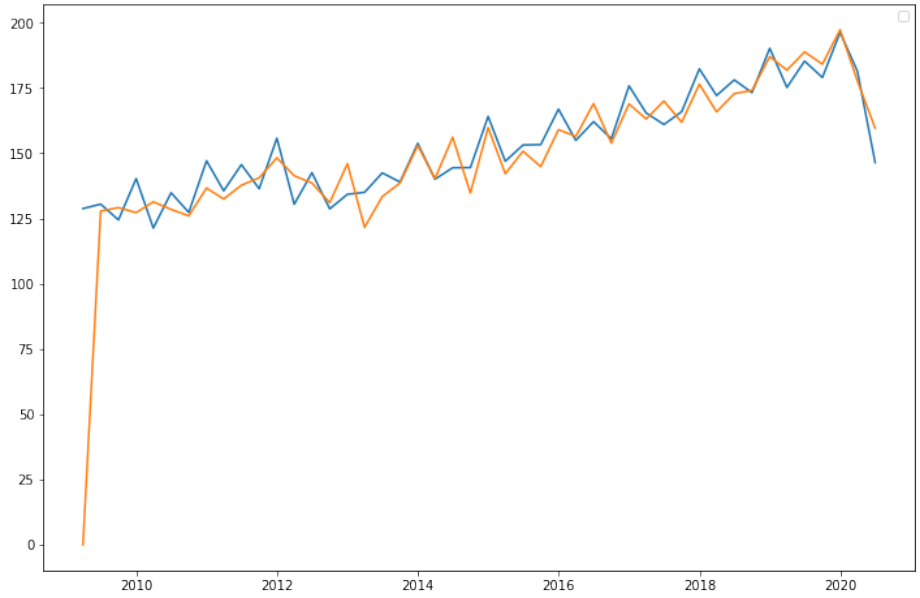

Рисунок 13 – Результат роботи моделі на тестових даних

Маємо побудовані моделі та брані параметри за допомогою яких можна проводити прогнозування майбутніх даних.

3.2 Прогнозування моделями математичної статистики та нейромереж

За допомогою попередньо побудованих моделей перейдемо до практичного використання наших побудованих моделей для прогнозування майбутніх значень податкових доходів.

3.2.1 Моделі ARIMA

Проводимо прогнозування моделлю ARIMA(1,1,4) щорічних даних та прогнозування методом $SARIMA(3,0,3)(14,0,12)_4$ квартальних даних, рисунки 14.1та 14.2 відповідно. Зеленим кольором наведено прогнозування, синім кольором навчальні дані ряду та оранжевим кольором оригінальний ряд, що прогнозувався.

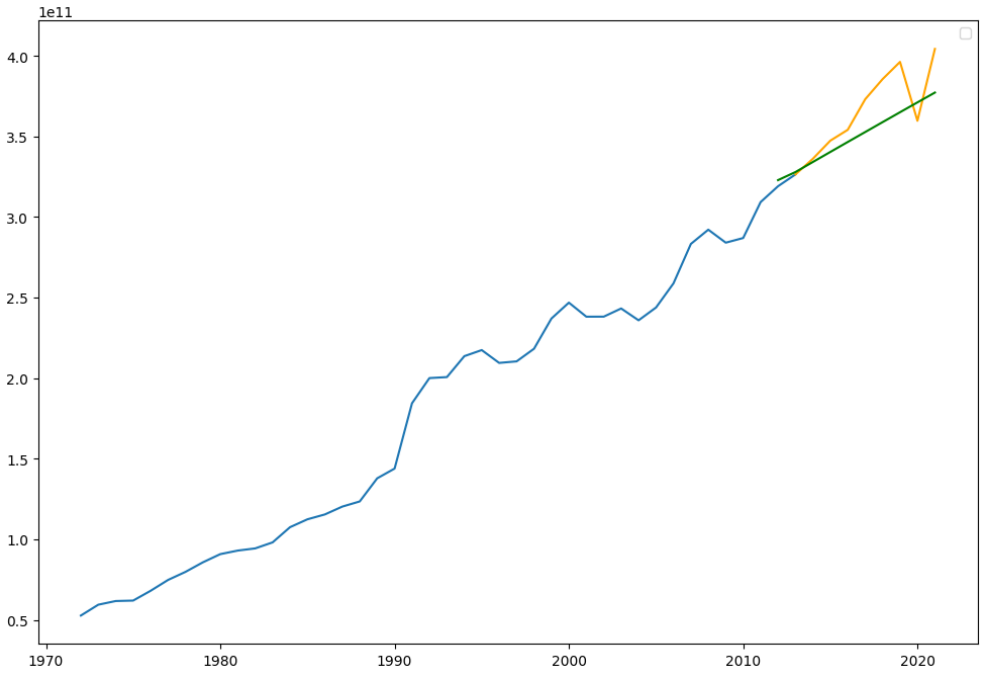

Рисунок 14.1 – Прогнозування щорічних податкових доходів моделлю авторегресійної інтегрованої рухомої середньої

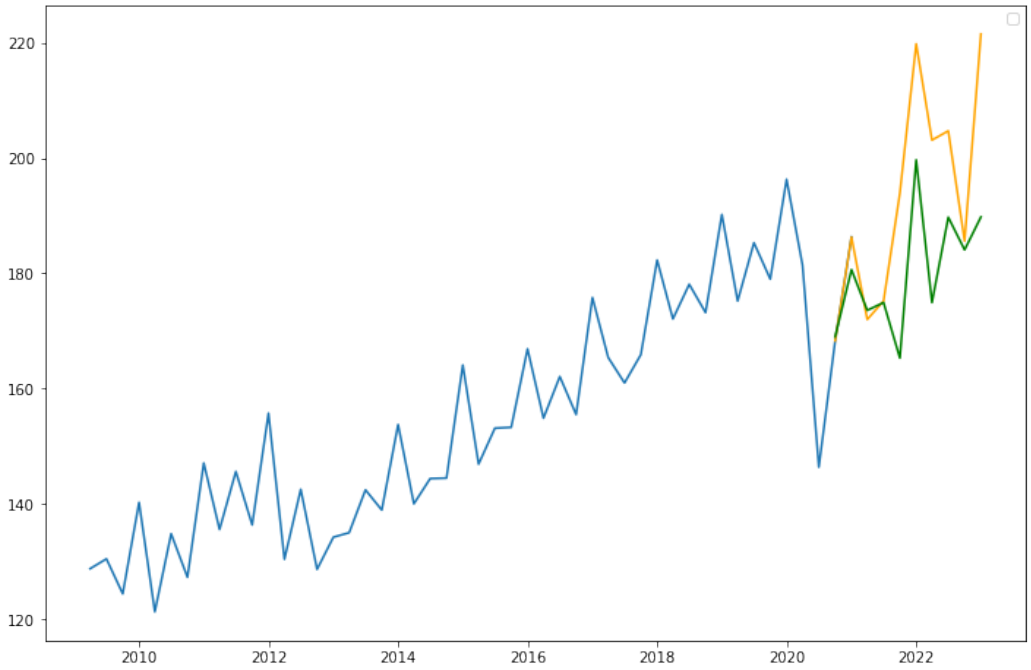

Рисунок 14.2 – Прогнозування квартальних податкових доходів моделлю сезонної авторегресійної інтегрованої рухомої середньої

3.2.2 Модель LSTM

Проводимо прогнозування за допомогою нейронних мереж, використовуючи модель двонаправленої довгої короткочасної пам’яті.
На рисунку 15.1. наведено прогнозування з використанням 60 епох(один повний прохід по тренувальному набору даних), червоним кольором наведений графік числового ряду, що був узятий для навчання, оранжевим кольором оригінальний ряд та зеленим прогноз.

На рисунку 15.2 прогноз квартальних податкових доходів з використанням 80 епох для навчання, так само, червоним кольором наведений графік числового ряду, що був узятий для навчання, оранжевим кольором оригінальний ряд та зеленим прогноз.

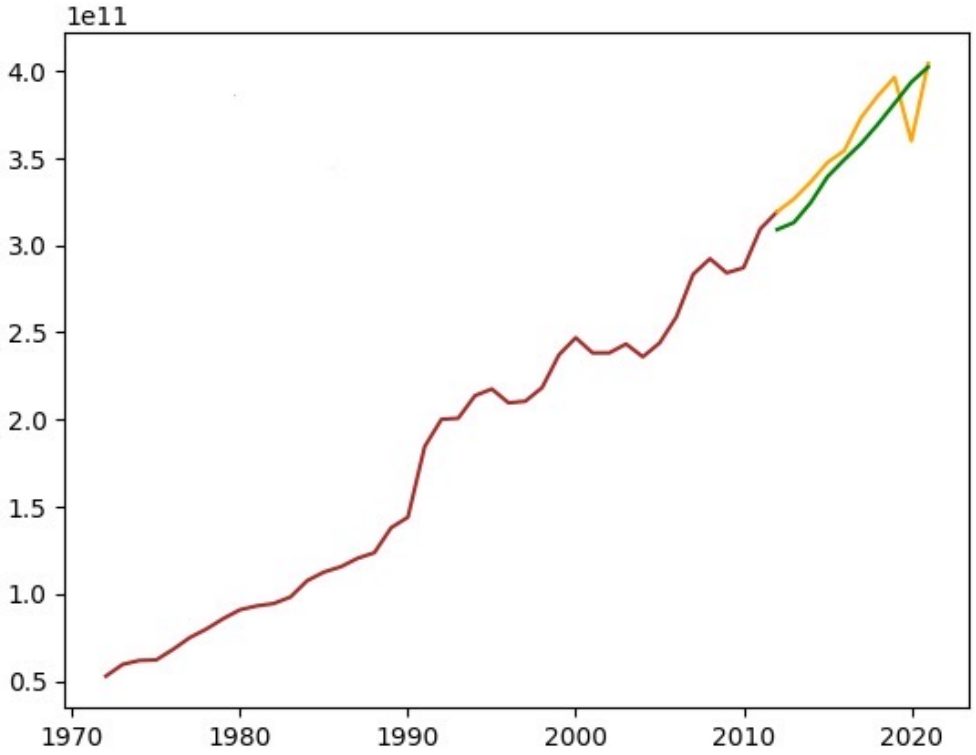

Рисунок 15.1 – Прогнозування щорічних податкових доходів моделлю двонаправленої довгої короткочасної пам'яті

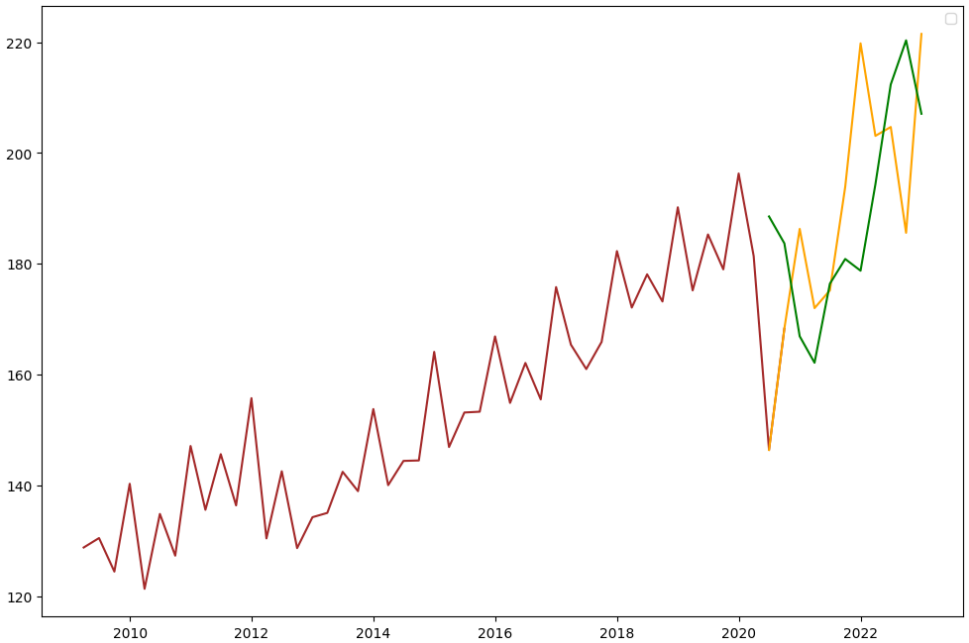

Рисунок 15.2 -– Прогнозування квартальних. податкових доходів моделлю двонаправленої довгої короткочасної пам'яті

3.3 Аналіз та порівняння моделей

Проведемо порівняння моделей.

Почнемо з порівняння прогнозів даних що були задані щорічно, на рисунку 16.1 маємо прогнозування методом ARIMA та на рисунку 16.2 прогноз побудовано за допомогою рекурентної нейронної мережі.

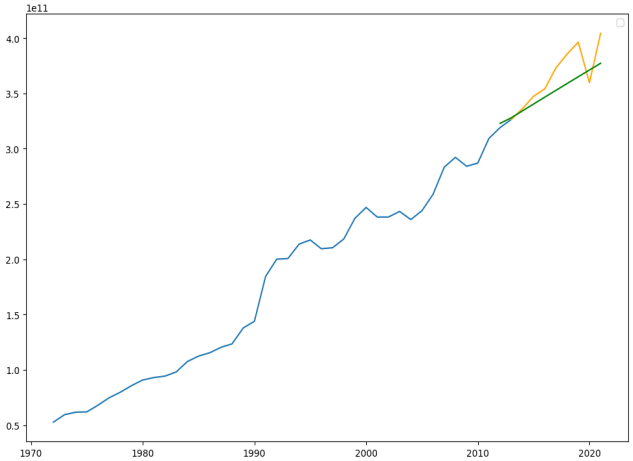

Рисунок 16.1 – Графік прогнозу податкових доходів, що задані щорічно методом математичної статистики

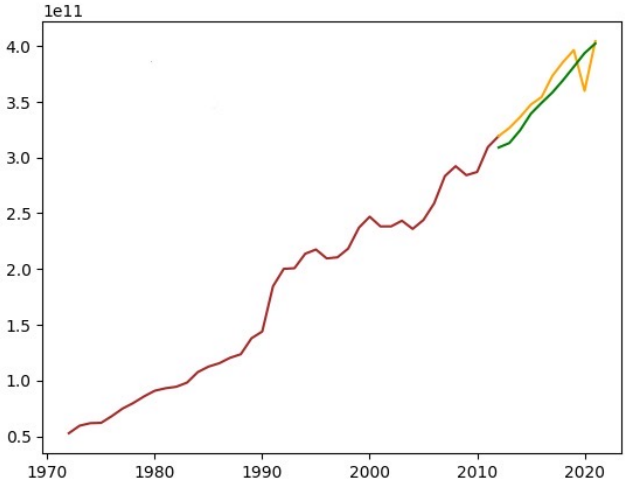

Рисунок 16.2 – Графік прогнозу податкових доходів, що задані щорічно методом нейронних мереж

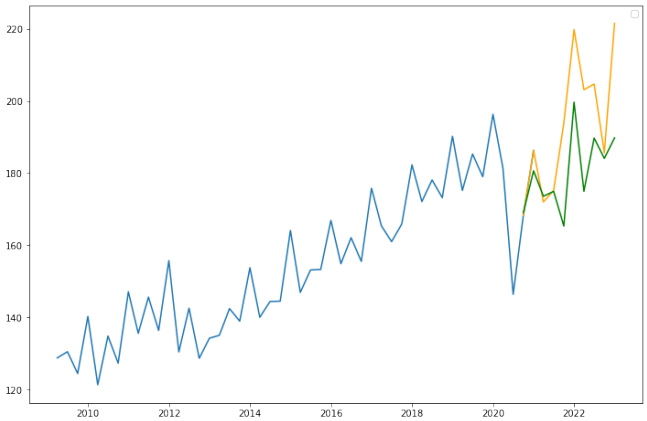

Рисунок 17.1 – Графік прогнозу податкових доходів, що задані квартально методом математичної статистики

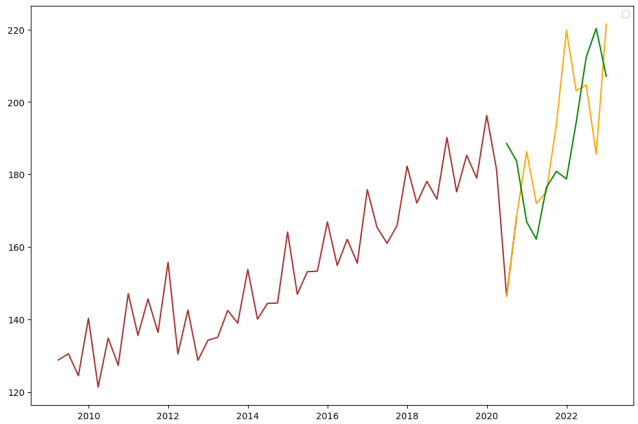

Рисунок 17.2 – Графік прогнозу податкових доходів, що. задані квартально методом нейронних мереж

ВИСНОВКИ

Для прогнозування часових рядів існують багато різних методів, методи математичної статистики є добре вивченими та ефективними для економічних даних, їх аналізу та прогнозування, у той час як нейромережі є розвиваючимся та сучасним напрямком аналізу, навчання та прогнозування, що є дуже актуальною темою.

Для застосування методів аналізу і прогнозування були обрані часові ряди податкових доходів, так як вони є складними для аналізу даними та є актуальними для прогнозування.

Під час аналізу даних виявилось, що щорічні дані, які були взяті зі Світового банку не мають достатньої сезонності за результатами аналізу методів, або ж мають надто складну сезонність яку не вдалось відслідкувати. Тому для розширення області аналізу та застосування методів було обрані дані двох різних типів: дані податкових доходів з 1972 по 2021рр., що були задані щорічно та з 2009 по 2022рр. поквартально. Аналіз показав, що дані занотовані щорічно мають сильний тренд та не проявляють сезонності на відміну від поквартальних даних, що є більш чутливими до змін, через частіше відслідковування.

Дані були розподілені на дві частини: навчальні дані, для тренування моделей (80%) та тестові дані для перевірки точності результатів (20%).

Для прогнозування даних методами математичної статистики був зроблений аналіз даних, для дослідження їх на стаціонарність, кількість потрібних диференцювань та для правильного вибору методу. Для методів ARIMA та SARIMA, що були обрані шляхом комплексного аналізу, були побудовані графіки, виконаний візуальний аналіз та побудовані різні експериментальні моделі для досягнення кращого результату у виборі параметрів.

Для рекурентних нейронних мереж став важливим та досить вирішальним вибір правильних параметрів, зважаючи на кількість тренувань, відсоток

перевірки та розподілення даних. Було застосоване візуальне та математичне порівняння результатів для використання різних параметрів та різного розподілення даних, для вибору кращих результатів.

Щодо порівняльного аналізу моделей математичної статистики та нейромереж, то метод LSTM рекурентних нейронних мереж виявився більш точним для щорічних даних, адже дані мали виражений тренд, який легко відслідкували нейромережі. Для квартальних даних були характериними і тренд, і сезонність, які стали важчими для роботи нейромереж, зважаючи на невелику кількість даних. Тому для сезонних даних квартальних податкових доходів найкраще спрацював метод SARIMA, тобто сезонне доповнення методу ARIMA. Метод був створений для відстеження сезонності та прогнозування з її використанням, що значно підвищило точність прогнозування.

Варто наголосити, що складність полягала у присутності в обох даних аномалій пов’язаних з COVID-19 у 2020 та 2021 роках.

Отже, всі методи продемонстрували хороші результати, для кращого прогнозування податкових доходів заданих квартально підійшов метод SARIMA, а для щорічних даних метод рекурентних нейромереж. Важливо додати, що при наявності більшої кількості даних податкових доходів усі методи мали би більшу точність та кращі результати прогнозування.

Для економічних даних, що є достатньо складними у прогнозуванні, варто використовувати методи математичної статистики та нейромереж у поєднанні, що допоможе створити модель подвійної дії, яка зменшить вплив викидів, що пов’язані з аномаліями та збільшить точність прогнозування.

ЛІТЕРАТУРА

[1] Світовий банк [Електронний ресурс] – Режим доступу до ресурсу: https://www.worldbank.org/en/home.

[2] Міністерство фінансів Німеччини [Електронний ресурс] – Режим доступу до ресурсу: bundesfinanzministerium.de.

[3] Jonathan D. Cryer. Time Series Analysis With Applications in R / Jonathan D. Cryer, Kung-Sik Chan. – New York, NY, 2008. – 508 c. – (Springer). – (Second Edition).

[4] Peter J. Brockwell. Time Series: Theory and Methods / Peter J. Brockwell, Richard A. Davis., 2009. – 580 c. – (Springer New York). – (Springer Series in Statistics).

[5] Klaus Neusser. Time Series Econometrics / Klaus Neusser., 2016. – 409 c. – (Springer). – (Springer Texts in Business and Economics).

[6] Dinesh C.S. Bisht. RECENT ADVANCES IN TIME SERIES FORECASTING / Dinesh C.S. Bisht, Mangey Ram., 2021. – 238 c. – (Mathematical engineering, manufacturing and management sciences).

[7] Peter J. Brockwell. Introduction to Time Series and Forecasting / Peter J. Brockwell, Richard A. Davis., 2016. – 425 c. – (Third Edition).

[8] John B. Guerard, Jr. Introduction to Financial Forecasting in Investment Analysis / John B. Guerard, Jr.. – Springer Science+Business Media New York 2013, 2013. – 236 c.

[9] Jaromír Vrbka. Using Artificial Neural Networks for Timeseries Smoothing and Forecasting / Jaromír Vrbka., 2021. – 189 c. – (Case Studies in Economics).

[10] Burcu Adıgüzel Mercangöz Editor. Handbook of Research on Emerging Theories, Models, and Applications of Financial Econometrics / Burcu Adıgüzel Mercangöz Editor., 2021. – 456 c.

[11] Recurrent Neural Networks for Short-Term Load Forecasting An Overview and Comparative Analysis / Filippo Maria Bianchi, Enrico Maiorino, Michael C. Kampffmeyer та ін.]., 2017. – (SpringerBriefs in Computer Science).

[12] Лекційні матеріали університету Buffalo [Електронний ресурс] – Режим доступу до ресурсу: https://cedar.buffalo.edu/~srihari/CSE676/.

[13] Рекурентні нейронні мережі [Електронний ресурс] // Факультет комп'ютерних наук, Стенфордський Університет – Режим доступу до ресурсу: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks.

[14] Günter Daniel Rey. Neuronale Netze / Günter Daniel Rey, Karl F. Wender., 2018. – 216 c. – (Hogrefe AG).

[15] Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling [Електронний ресурс] // Cornell University. – 2014. – Режим доступу до ресурсу: https://arxiv.org/abs/1412.3555.

[16] Ian Goodfellow. Deep Learning / Ian Goodfellow, Yoshua Bengio, Aaron Courville., 2016. – 800 c. – (The MIT Press).

[17] Особистий інтернет ресурс [Електронний ресурс] – Режим доступу до ресурсу: https://github.com/Polinden.

ДОДАТОК А

Dep. Variable:	y	**No. Observations:**	40
Model:	ARIMA(1, 1, 5)	**Log Likelihood**	-950.040
Date:	Sun, 04 Jun 2023	**AIC**	1914.080
Time:	14:51:46	**BIC**	1925.725
Sample:	0	**HQIC**	1918.259
	- 40		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9989	0.026	38.324	0.000	0.948	1.050
ma.L1	-0.8235	0.411	-2.002	0.045	-1.630	-0.017
ma.L2	-0.3287	0.404	-0.814	0.415	-1.120	0.462
ma.L3	0.0533	0.400	0.133	0.894	-0.730	0.837
ma.L4	0.0552	0.428	0.129	0.898	-0.784	0.895
ma.L5	0.0610	0.365	0.167	0.867	-0.655	0.777
sigma2	1.097e+20	2.08e-20	5.26e+39	0.000	1.1e+20	1.1e+20

Ljung-Box (L1) (Q):	0.13	**Jarque-Bera (JB):**	56.27
Prob(Q):	0.72	**Prob(JB):**	0.00
Heteroskedasticity (H):	9.47	**Skew:**	1.54
Prob(H) (two-sided):	0.00	**Kurtosis:**	8.01

Рисунок А.1

```
Ljung-Box test:
     lb_stat  lb_pvalue
10  4.883619   0.898815
ADF Test:
ADF Statistic: -8.442595211986989
p-value: 1.7468825828068823e-13
```

Рисунок А.2

```
Ljung-Box test:
    lb_stat  lb_pvalue
10  1.07085   0.999765
ADF Test:
ADF Statistic: -19.395838511618933
p-value: 0.0
```

Рисунок А.3

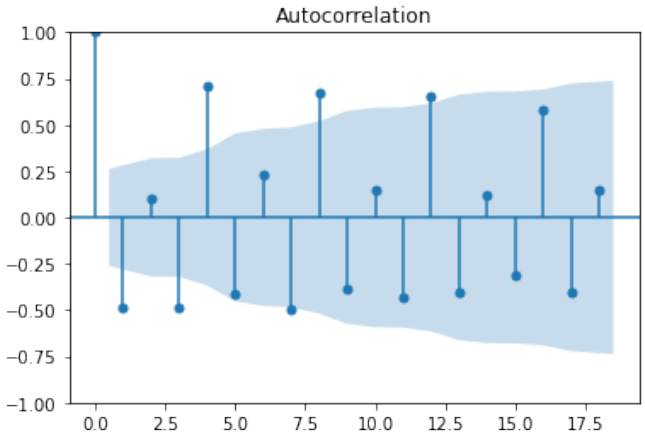

Рисунок А.4 – Графік автокореляційної функції сезонної компоненти податкових доходів Німеччини по кварталам

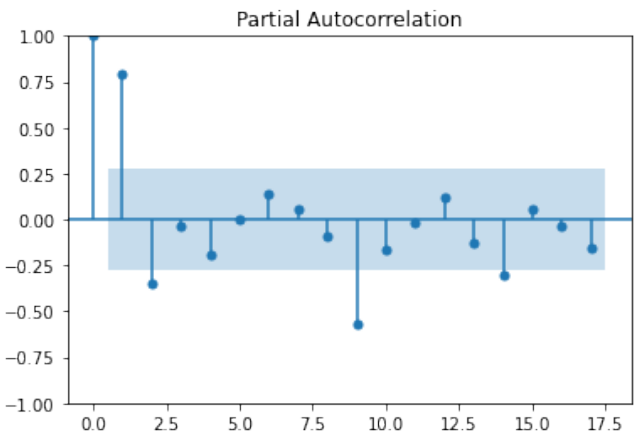

Рисунок А.5 – Графік автокореляційної функції сезонної компоненти податкових доходів Німеччини по кварталам

UNICHECK

Ім'я користувача:
Оноцький В'ячеслав ФКомпНаук

ID перевірки:
1015590420

Дата перевірки:
13.06.2023 18:58:16 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
13.06.2023 18:58:45 EEST

ID користувача:
100002816

Назва документа: **ЛипницькаПолінаДенисівна**

Кількість сторінок: **48** Кількість слів: **7131** Кількість символів: **51279** Розмір файлу: **1.04 MB** ID файлу: **1015239656**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

3.73%
Схожість

Найбільша схожість: **1.22%** з Інтернет-джерелом (https://powcoder.com/2021/01/23/%E7%A8%8B%E5%BA%8F%E4%BB%..

2.41% Джерела з Інтернету	84	Сторінка 50
1.67% Джерела з Бібліотеки	55	Сторінка 50

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	280
Підозріле форматування	12 сторінок

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

СИСТЕМА ЗАПОБІГАННЯ ТА ВИЯВЛЕННЯ АКАДЕМІЧНОГО ПЛАГІАТУ

Довідка про оригінальність кваліфікаційної роботи за освітнім рівнем бакалавр

Експертна оцінка роботи науковим керівником :

Робота студенки 4-го курсу Липницької Поліни Денисівни «Аналіз часових рядів для прогнозування податкових доходів засобами математичної статистики та нейромереж» виконана самостійно, при цьому обсяг цитувань та запозичень становить 3.73% та не перевищує норму.

Науковий керівник: _____ Тимошенко А.А.
(підпис) (ПІБ)

Оператор: _____ Оноцький В.В.
(підпис) (ПІБ)

ВІДГУК
на випускну кваліфікаційну роботу бакалавра
«Аналіз часових рядів для прогнозування податкових доходів засобами математичної статистики та нейромереж»
студентки 4-го курсу кафедри обчислювальної математики
факультету комп'ютерних наук та кібернетики
Київського національного університету імені Тараса Шевченка
Липницької Поліни Денисівни

В даній роботі розглянуто та застосовано методи математичної статистики та нейронних мереж для аналізу та прогнозування часових рядів. Ці методи застосовано для прогнозування податкових доходів країн, використовуючи: метод аналізу даних ARIMA (Autoregressive Integrated Moving Average), сезонний варіант цього методу SARIMA, модель LSTM рекурентних нейронних мереж. При цьому наведено короткий опис використаних підходів та графічне зображення справжніх даних у порівнянні з моделлю.

Розглянуті підходи є корисними для моделювання процесу та здійснення короткострокового прогнозу. Використання кількох альтернативних способів аналізу дає додаткові можливості передбачення з урахуванням сезонних явищ.

Робота виконана самостійно, отримані результати мають практичну цінність. Роботу можу оцінити на «відмінно», студентка заслуговує на присвоєння відповідного освітньо-кваліфікаційного рівня.

Асистент кафедри обчислювальної
математики факультету комп’ютерних
наук та кібернетики
Київського національного
університету
імені Тараса Шевченка,
доктор філософії

Андрій
ТИМОШЕНКО

РЕЦЕНЗІЯ

на випускну кваліфікаційну роботу бакалавра
«Аналіз часових рядів для прогнозування податкових доходів засобами математичної статистики та нейромереж»
студентки 4-го курсу кафедри обчислювальної математики
факультету комп'ютерних наук та кібернетики
Київського національного університету імені Тараса Шевченка
Липницької Поліни Денисівни

У роботі розглянуто задачу моделювання та прогнозування даних методами математичної статистики ARIMA та його доповнення з урахуванням сезонності SARIMA, метод LSTM рекурентних нейронних мереж, виконано пошук тренду, який легко відслідкували нейроні мережі. Для сезонних даних квартальних податкових доходів найкраще спрацював метод SARIMA.

У роботі коротко описано використані методи, причини їх вибору та виявлені переваги. Графічна демонстрація сприяє легкому порівнянню результатів.

Студентка виконала якісне дослідження та отримала корисні результати. Вважаю, що вона заслуговує на оцінку «відмінно», а також на присвоєння кваліфікації бакалавра.

Професор кафедри обчислювальної
математики факультету комп'ютерних
наук та кібернетики
Київського національного університету
імені Тараса Шевченка,
доктор фізико-математичних наук

Дмитро КЛЮШИН